

Title:

Response to: A new method for estimating animal abundance with two sources of data in capture-recapture studies

Author Information:

Simon Bonner
Department of Statistics
University of Kentucky
Rm. 311 Multidisciplinary Science Building
725 Rose Street
Lexington, KY 40536-0082
USA

Email: simon.bonner@uky.edu

Phone: 859-257-4950

Fax: 859-323-1973

Running Head:

Response to Madon et al. (2011).

Word Count:

This manuscript contains approximately 2100 words excluding displayed equations.

Summary

1. Mark-recapture studies that rely on multiple marks to identify individuals pose modeling challenges if the marks for each individual are not always linked. If an individual with unlinked marks is encountered on two occasions and different marks are observed then it will appear that two different individuals were captured. Failing to account for these missed matches will produce incorrect inference.
2. Madon et al. (2011) proposes a modification of the Jolly-Seber estimator for such data computed by adjusting the observed counts of individuals first captured, recaptured, or not captured but known to be alive on each occasion. The adjustment involves multiplying each of these counts by a constant factor, I_{id} , intended to correct for double counting of individuals and constrained between 0 and 1. Results of a simulation study provided in Madon et al. (2011) show that the proposed estimator is almost unbiased, but its uncertainty is underestimated and the true coverage of confidence intervals is consistently below the nominal value.
3. I compute separate adjustment factors for each of the counts and show 1) that a constant adjustment is not appropriate and 2) that the theoretical adjustment factor is sometimes greater than 1. I believe that the use of a single adjustment factor between 0 and 1 is what causes the uncertainty to be underestimated and that complete models of the observation process are required to obtain valid results.

Keywords

Mark-recapture; Genetic markers; Multiple marks; Non-invasive marks; Photo identification

Natural or non-invasive marks, including skin patterns and genetic markers, allow individuals to be identified in mark-recapture studies without applying man-made tags. However, data from natural marks also present novel statistical challenges. One challenge that has seen little discussion concerns inference from data that include multiple natural marks that cannot be linked unless they are observed together. This occurs if, for example, individuals are identified from skin patterns on different parts of the body or separately from both photo-identification of skin patterns and from genetic markers. Marks for one individual may not be linked, meaning that it is not always possible to determine if, say, a photo taken on one occasion and a genetic sample collected on another occasion represent the same individual or two different individuals. Analyzing the data without accounting for the possibility of missed matches will erroneously inflate the number of individuals encountered and create dependence between the histories, violating a key assumption of most mark-recapture models. Madon et al. (2011) proposes a method for estimating abundance from such data using an adjusted version of the Jolly-Seber (JS) estimator. However, the results of the simulation study provided in Madon et al. (2011) suggest that the method does not perform well. Though the adjusted JS estimator is approximately unbiased, its sampling variance is consistently underestimated particularly when capture probabilities are high. The average coverage probability of the 95% confidence intervals is .84 when data are simulated from the true model with capture probability .05 but only .45 when the capture probability is .80 (Madon et al., 2011, Table 2). That the coverage probabilities are worse when capture probabilities are higher – i.e. when the data contain more information – is disturbing and 95% confidence intervals with coverages as low as 45% will provide misleading information for biologists trying to understand or manage a

population. I believe that these results stem from errors in the derivation of the adjusted JS estimator which I discuss below.

The specific objective of Madon et al. (2011) is to estimate the number of humpback whales breeding around the island of New Caledonia each year between 1996 and 2001. Data come from a study in which whales were identified from both photo-identification of skin patterns (mark 1) and collection of DNA samples (mark 2). Sampling occurred from July to September and it is assumed that individuals can only be encountered once each year. Each year, one of four events may occur:

- 0) the individual is not encountered,
- 1) the individual is encountered and only mark 1 is seen,
- 2) the individual is encountered and only mark 2 is seen,
- 3) the individual is encountered and both marks are seen together.

The key assumption is that the two marks for an individual can be matched only if both are seen together in at least one year. If this occurs then I say that the individual's marks are linked. If it does not then I will say that the individual's marks are not linked. If an individual's marks are linked (i.e., event 3 occurs in the encounter history) then its true capture history can be known with certainty. However, if an individual's marks are not linked and both of its marks were observed on separate occasions then it will contribute two separate histories to the observed data. For example, an individual with the true history 01120 will contribute the histories 01100 and 00020 to the observed data – exactly as if two different individuals had been encountered.

The primary challenge with modeling such data is not to provide point estimates of the demographic parameters (e.g., the population size) but rather to compute appropriate measures of precision. Suppose that individuals can be identified by one of p marks. A simple estimator of population size combining the data from all marks is given by obtaining separate estimates for each mark, $\hat{N}_1, \dots, \hat{N}_p$, and then computing the weighted average

$$\hat{N} = \sum_{m=1}^p w_m \hat{N}_m$$

for some set of positive weights such that $\sum_{m=1}^p w_m = 1$. The bias of this estimator will be the weighted average of the bias of the separate estimators, and \hat{N} will be unbiased/consistent if each \hat{N}_m is unbiased/consistent. However, the standard error of \hat{N} is difficult to compute because the separate estimates $\hat{N}_1, \dots, \hat{N}_p$ are dependent but their correlations are unknown. Wilson et al. (1999) describes an application of this method for two marks using inverse variance weights so that $w_m \propto 1/\widehat{\text{Var}}(\hat{N}_m)$, $m = 1, 2$. The variance of \hat{N} is then approximated by treating $\hat{N}_1, \dots, \hat{N}_p$ as independent so that $\widehat{\text{Var}}(\hat{N}) = \sum_{m=1}^p w_m^2 \widehat{\text{Var}}(\hat{N}_m)$, but this is not appropriate and $\widehat{\text{Var}}(\hat{N})$ will be biased low.

To account for the possibility that some whales contribute multiple encounter histories to the New Caledonia humpback data, Madon et al. (2011) proposes an adjustment of the JS estimator of abundance. Specifically, Madon et al. (2011) builds on the bias-corrected version of the JS estimator, assuming no losses on capture, given by

$$\hat{N}_i = \frac{m_i + U_i + 1}{m_i + 1} \cdot \frac{m_i + U_i + 1}{r_i + 1} \cdot \frac{z_i}{m_i}$$

where

- U_i is the number of individuals first encountered on occasion i ,
- m_i is the number of individuals encountered both before and on occasion i ,
- r_i is the number of individuals encountered both on and after occasion i , and
- z_i is the number of individuals encountered both before and after but not on occasion i .

Uncertainty in matching observed histories with different marks means that these counts cannot be known, and so Madon et al. (2011) attempts to adjust these values before computing \hat{N}_i .

Consider m_i . This quantity cannot be observed because it is not always possible to know whether or not an individual encountered on occasion i was previously encountered. For example, if an individual's true history was 01200 then it is truly encountered both before and on occasion 3, but this cannot be known because its marks are not linked. This individual will be mistakenly excluded from the observed count of m_2 . To account for such errors, Madon et al. (2011, eqn 4) replaces m_i with

$$m_i \approx m_{i,3}^* + (m_{i,1}^* + m_{i,2}^*)I_{id} \tag{1}$$

where

- $m_{i,1}^*$ represents the number of individuals for which mark 1 was seen prior to occasion i and mark 1 was seen on occasion i and whose marks are not linked,

- $m_{i,2}^*$ represents the number of individuals for which mark 2 was seen prior to occasion i and mark 2 was seen on occasion i and whose marks are not linked, and
- $m_{i,3}^*$ represents the number of individuals encountered before and on occasion i whose marks are linked.

The raw sum $m_{i,1}^* + m_{i,2}^* + m_{i,3}^*$ is the number of individuals known to be captured both before and on occasion i , and I_{id} is an adjustment factor intended to account for the difference between this value and the true number of individuals recaptured. It is described as the “Probability of true identity” (Madon et al., 2011, pg. 3) and assumed to be between 0 and 1. The same adjustment is also used for U_i , r_i , and z_i so that (Madon et al., 2011, eqns 5, 6, and 7)

$$U_i \approx U_{i,3}^* + (U_{i,1}^* + U_{i,2}^*)I_{id} \quad (2)$$

$$r_i \approx r_{i,3}^* + (r_{i,1}^* + r_{i,2}^*)I_{id} \quad (3)$$

$$z_i \approx z_{i,3}^* + (z_{i,1}^* + z_{i,2}^*)I_{id} \quad (4)$$

where $U_{i,j}^*$, $r_{i,j}^*$, and $z_{i,j}^*$ are defined analogously to $m_{i,j}^*$, $j = 1, 2, 3$. My concerns with these adjustments are 1) that all counts are adjusted by the same factor and 2) that I_{id} is assumed to be less than 1. I show below that the correct adjustment factors for the different counts are not the same and that some of the observed counts may underestimate the true counts so that the correct adjustment factor may be greater than 1.

To illustrate my concerns, I derive adjustment factors for m_2 , U_2 , r_2 , and z_2 for the specific case of an experiment with 3 occasions. In the following discussion, I use

$n(\omega_1, \omega_2, \omega_3)$ to denote the number of individuals whose true encounter history is $\omega = (\omega_1, \omega_2, \omega_3)$. Events on each occasion are represented using the notation on page 4. For convenience, I use -3 to represent any event except 3; so that, e.g., $n(-3, 1, 1) = n(0, 1, 1) + n(1, 1, 1) + n(2, 1, 1)$. Note that some of these counts cannot be observed. For example, $n(0, 1, -3)$ and $n(2, 1, -3)$ cannot be distinguished even though their sum is known.

Consider m_2 . The true number of individuals encountered both before occasion 2 and on occasion 2 is

$$m_2 = n(1, 1, -3) + n(1, 2, -3) + n(2, 1, -3) + n(2, 2, -3) + m_{2,3}^*. \quad (5)$$

From the definitions above

$$m_{2,1}^* = n(1, 1, -3) \text{ and } m_{2,2}^* = n(2, 2, -3). \quad (6)$$

Substituting (5) and (6) into (1) and solving for I_{id} then yields

$$\begin{aligned} I_{id}^m &= \frac{n(1, 1, -3) + n(1, 2, -3) + n(2, 1, -3) + n(2, 2, -3)}{n(1, 1, -3) + n(2, 2, -3)} \\ &= 1 + \frac{n(1, 2, -3) + n(2, 1, -3)}{n(1, 1, -3) + n(2, 2, -3)} \end{aligned} \quad (7)$$

Note that $I_{id}^m \geq 1$. In general, errors in counting the number of individuals encountered both before and on occasion i occur because different marks were seen and the encounters cannot be linked. This means that the observed value underestimates the true value of m_i and should be adjusted upward.

Next consider the count of individuals first encountered on occasion 2, U_2 . The true value is

$$U_2 = n(0, 1, -3) + n(0, 2, -3) + U_{2,3}^* \quad (8)$$

whereas

$$U_{2,1}^* = n(0, 1, -3) + n(2, 1, -3) \text{ and } U_{2,2}^* = n(0, 2, -3) + n(1, 2, -3) \quad (9)$$

Substituting (8) and (9) into (2) then yields

$$\begin{aligned} I_{id}^U &= \frac{n(0, 1, -3) + n(0, 2, -3)}{n(0, 1, -3) + n(2, 1, -3) + n(0, 2, -3) + n(1, 2, -3)} \\ &= 1 - \frac{n(2, 1, -3) + n(1, 2, -3)}{n(0, 1, -3) + n(2, 1, -3) + n(0, 2, -3) + n(1, 2, -3)} \end{aligned} \quad (10)$$

Clearly $I_{id}^U \neq I_{id}^m$ and $I_{id}^U \leq 1$. In general, errors in counting the number of individuals first encountered on occasion i occur because some individuals were actually encountered before but the marks seen before and on occasion i were not the same and are not linked. This leads to overcounting of the number of individuals first encountered on occasion i so that the observed count should be reduced.

Theoretical correction factors for r_2 and z_2 can be computed in the same way. The adjustment factor for r_2 is

$$\begin{aligned} I_{id}^r &= \frac{n(-3, 1, 1) + n(-3, 1, 2) + n(-3, 2, 1) + n(-3, 2, 2)}{n(-3, 1, 1) + n(-3, 2, 2)} \\ &= 1 + \frac{n(-3, 1, 2) + n(-3, 2, 1)}{n(-3, 1, 1) + n(-3, 2, 2)} \end{aligned} \quad (11)$$

Then $I_{id}^r \neq I_{id}^m$, $I_{id}^r \neq I_{id}^U$, and again $I_{id}^r \geq 1$. Generally, the number of individuals encountered on and after occasion i is under-counted because some individuals with

unlinked marks are missed. Finally, the adjustment factor for z_2 is

$$\begin{aligned}
 I_{id}^z &= \frac{n(1, 0, 1) + n(1, 0, 2) + n(2, 0, 1) + n(2, 0, 2)}{n(1, 0, 1) + n(2, 0, 2) + n(1, 2, 1) + n(2, 1, 2)} \\
 &= 1 + \frac{[n(1, 0, 2) + n(2, 0, 1)] - [n(1, 2, 1) + n(2, 1, 2)]}{n(1, 0, 1) + n(2, 0, 2) + n(1, 2, 1) + n(2, 1, 2)}.
 \end{aligned}
 \tag{12}$$

This value may be either greater or less than 1 depending on whether the number of individuals mistakenly included in z_2 and doubly counted, $n(1, 2, 1) + n(2, 1, 2)$, is bigger or smaller than the number of individuals mistakenly excluded from z_2 , $n(1, 0, 2) + n(2, 0, 1)$.

Equations (7), (10), (11), and (12) for a study with 3 occasions make it clear that the proper adjustment factors for m_2 , U_2 , r_2 , and z_2 are not the same and are not all constrained to $(0, 1)$. The expressions for studies with more capture occasions are more complicated but the same results hold in general. Given these results, it is not surprising that the variance of \hat{N}_i is underestimated when the adjustment factors are all assumed to be equal and less than 1. Of course, the correct factors cannot be computed explicitly because they depend on counts of individuals with unlinked marks that cannot be observed like $n(2, 1, -3)$, $n(-3, 1, 2)$, and $n(1, 0, 2)$. To compute valid adjustment factors one would need to estimate these quantities by modeling the observation process, and so I recommend a model based approach. Maximum likelihood estimation provides one solution but is complicated because the exact likelihood can only be computed by summing over all possible configurations of the true histories, and this set may be very complicated. An alternative is to apply a Bayesian complete-data likelihood approach treating the counts of the true histories as a latent-multinomial process, as in Link et al. (2010). Bonner and Holmberg (2012) examines this approach using Markov chain Monte Carlo to sample

simultaneously from the posterior distribution of both the model parameters and the configuration of the true encounter histories. Advantages of this approach are that it provides valid interval estimates with nominal coverage, and also produces inference about the population dynamics (e.g., survival probabilities, recruitment probabilities, and population growth rates) as well as population size. Furthermore, the model of Bonner and Holmberg (2012) allows for individuals to be encountered multiple times during each capture occasion. Though this is not a problem with the method of Madon et al. (2011) if capture occasions are instantaneous, the assumption that humpback whales are encountered only once each year seems restrictive given that sampling occurs for 3 months of the year.

References

- Bonner, S. J. and Holmberg, J. (2012). Mark-recapture with multiple non-invasive marks. arXiv 1208.4285. Submitted to *Biometrics* in August, 2012.
- Link, W. a., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: Analysis of mark-recapture data with misidentification. *Biometrics*, 66(1):178–185.
- Madon, B., Gimenez, O., McArdle, B., Scott Baker, C., and Garrigue, C. (2011). A new method for estimating animal abundance with two sources of data in capture-recapture studies. *Methods in Ecology and Evolution*, 2(4):390–400.

Wilson, B., Hammond, P. S., and Thompson, P. M. (1999). Estimating size and assessing trends in a coastal bottlenose dolphin population. *Ecological Applications*, 9(1):288–300.